Measurement Options for Development of Sustainable Development Goal Indicator 4.2.1

Memo, Global Alliance to Monitor Learning, Taskforce on Target 4.2

Prepared by Hirokazu Yoshikawa, Abbie Raikes, and Alice Wuermli [1]

June 2017 DRAFT

In this memo we describe options for developing a measurement strategy at the global level for Target 4.2, and specifically Indicator 4.2.1 ("Proportion of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being, by sex"). Goals for such a strategy include development of a measure that is appropriate for cross-country comparison of levels of development in each of these major domains (health, learning and its subdomains, and psychosocial well-being); feasible to use in national monitoring and evaluation, in terms of cost and human capital resources (e.g., training intensity and ease in achieving reliability); vertically equitable in its units[2] across the span of birth through 60 months; and psychometrically sound in reliability and validity from standpoints of both classical test theory and other approaches such as IRT.

Our memo has three parts: 1) Three options for a global measurement strategy; 2) weighing options at the intersection of validity and use; and 3) options for next steps.

A. **Three Options for A Global Measurement Strategy**.

Option 1. *An existing measure could be chosen without adaptation as a single global measure for Indicator 4.2.1.*

The SDGs whenever possible aim to achieve global indicators that are comparable across countries. In contrast to the MDGs, the countries in question include all UN member nations, and thus cut across high-, middle- and low-income countries. The choice to select a single global indicator for 4.2.1 is thus quite formidable; and setting a goal by 2030 of a single assessment, drawn from previous learnings across other measures that have been analyzed for validity across multiple countries, is important to weight carefully.

What domains of child development should such a measure assess? The language for Indicator 4.2.1 reflects a global consensus in the field of early childhood development regarding the

[1] Yoshikawa and Wuermli, NYU Global TIES for Children Center; Raikes, University of Nebraska Medical Center. We thank Elise Legault, Baela Raza Jamil, select members of the Taskforce on Target 4.2. of GAML, and Ivelina Borisova for comments on previous drafts.

[2] I.e., for a particular construct, a unit at one point of the scale's distribution is comparable to a unit at another point of the scale's distribution.

multi-domain nature of development in the first years of life. Domains of physical, cognitive, language, numeracy, and socio-emotional development are typically inter-related, yet distinct, within the age range covered (birth through 60 months). Although great variability occurs in the nature of behaviors and skills in these overall domains of development, both within and across countries, the consensus concerning the meaningfulness of these domains in contexts of national ECD policy and planning has been shown across multiple regions and nations (e.g., Kagan & Britto, 2005).

The inclusion of multiple domains including physical, cognitive, learning (e.g., language and numeracy) and socio-emotional domains represents a relatively strong consensus in measures that have recently been assessed across multiple nations (Raikes & Anderson, 2017). These include instruments based on adult / caregiver report, such as the EDI (Janus & Offord, 2007), the CREDI (McCoy et al., 2017) and the UNICEF Early Childhood Development Index (Bornstein et al., 2012; McCoy et al., 2016); as well as instruments that directly assess children, such as the PRIDI (Verdisco, Cueto, & Thompson, 2016), IDELA (Wolf et al., 2017), EAP-ECDS (Rao et al., 2014), the MELQO MODEL measure (UNESCO, 2017), and others.

Despite this range of recent efforts to measure, in coordinated fashion, multiple domains of early childhood development, currently no consensus measure exists for Indicator 4.2.1 that is measured across a large number of countries (across, e.g., low-, middle- and high-income countries) *and* meets other criteria for a Tier I indicator of the UN Statistical Commission (2016, 2017). Thus, although Option 1 would be ideal if all conditions were met for feasibility, relevance and validity across countries, in the current context of the SDG indicators, there is no alternative that meets these criteria. As discussed below, Option 3 would work towards creating such a single criterion measure in the future.

Option 2. *Use an existing common set of items or identify a set of anchor items to integrate into national and regional assessments.*

At present, there are a range of measures that have been developed and tested within countries and regions. An overview of these measures appears in the first background paper (Anderson & Raikes, 2017). Many countries have also expressed the desire to build (either by adapting or creating new) nationally-specific measures to promote ongoing monitoring of child development in a manner that is aligned with national standards and cultural expectations. Below we present two ideas on how common item sets or anchor items could be used in global measurement.

Common Outcome Sets. In various fields, the use of Common Outcome Sets (COS's) has been implemented to establish common sets of measures or items across a set of evaluation or other research studies (e.g., Gershon et al., 2013; Schmitt et al., 2015; Williamson et al., 2012). Across these initiatives, a typical multi-phase process includes the following. First, a consensus group of experts and practitioners / policy leaders is brought together to establish agreement on the constructs that will constitute a measurement domain. Second, criteria for measures that may be considered as candidates to contribute items or entire scales /

assessments are agreed upon. Third, an inventory of measures meeting these criteria is assembled, and common items or tasks are identified. Fourth, depending on the initiative, a single "consensus" measure may be developed (some aspects of which may be newly developed, with others drawn from existing measures). Fifth, phases of pilot testing, psychometric analyses, and revision may occur iteratively until a final consensus measure is agreed upon. Finally, a measure and its guidelines for administration may be disseminated to a wide range of potential users, with continued input and refinement as the measure enters general use.

In the area of Indicator 4.2.1, a recent initiative to develop a common set of items was carried out as part of the Measuring Early Learning and Quality Outcomes project (MELQO; UNESCO, 2017). MELQO began with the intent of clarifying if one measure would be sufficient for measurement in all countries (Option 1), but moved quickly in the direction of Option 2. Option 2, finding a common item set, was desirable for two main reasons: 1) because it would allow countries to build on existing measures that were already developed and validated in each region; and 2) it would allow a greater deal of flexibility to add more culturally-responsive, nationally-specific items that are not possible to include when relying on only one measure. Ultimately the common item set was distilled into a single new measure named the MODEL, covering developmental domains of social, cognitive, language and literacy, numeracy, and executive function. Cross-country analyses are underway on this measure (Raikes et al. 2017).

*Measures harmonization***.** Multiple existing measures may be harmonized using some form of standardization to allow for scoring on a common scale. Two approaches are outlined below (but other may be relevant): 1) crosswalk samples; and 2) identification of anchor items. The process of harmonizing across measures is typically done through identifying common items that can help link the different assessments ("anchor items") (Chan et al., 2015), and/or by administering multiple measures on the same sample to synchronize measurements and establish the basis for comparing children's learning and development on the same scale, but with data collected through different measures ("crosswalk samples"). To investigate the feasibility of this option, either multiple data sets with a set of common items are needed ("anchor items") or multiple measures must be administered to the same children ("crosswalk sample"), so that calibration across different measures can take place. For example, anchor items could be created from those measures used in multiple countries where certain items have shown evidence of cross-country invariance. It would then be necessary to ensure that this scale works in similar ways across countries, a related but distinct step in building international comparability.

Crosswalk samples. Crosswalk samples (single samples that incorporate assessment of multiple instruments) are useful for facilitating decisions about what to include and exclude from particular instruments in developing a single set of common items or reduced consensus assessment. This is because in a single sample, multiple alternative measures are assessed, allowing for direct calculations of correlations among measures, differences in predictive or other forms of validity that do not confound sample with measure. This approach has been used recently in a study that aimed to harmonize multiple measures of depression and subjective health among older adults (Gatz et al., 2015).

Anchor items. A challenge is how to identify anchor items in the absence of any universal measure or indicator. For example, a recent effort in the United States to harmonize state level standardized assessments relied on an anchor measure that is administered across the country, namely the National Assessment of Educational Progress (NAEP). Based on the distribution of districts on that national assessment, a standardization procedure was used to link the state-level assessments (Reardon, Kalogrides, & Ho, 2016). In the absence of a criterion or audit test such as the NAEP in the U.S. example, a mix of conceptual and traditional empirical criteria from both classical test theory and other approaches such as item-response analysis may be utilized. However, variation in task requirements, item language, assessors across data sets, order of items, and response categories all create daunting differences in sources of measurement error even when considered within domain (e.g., language or numeracy).

Option 3. *Create a new universal "criterion" scale of child development against which many other possible measures could be placed*

The development of a new universal criterion scale could proceed following established procedures somewhat similar to the ones that led to the MELQO, IDELA or regional measures, but with learnings synthesized from all of them. It could start with the two leading initiatives in this field, which are the UNICEF ECDI (longstanding, for 3-6 year olds) and the WHO consortium instrument for 0-3 year olds (more recently developed). The first step of pooling items from measures used in multiple countries, categorizing them by outcome domain, and compiling information on validity studies, samples, and countries, was also recently completed in the first phase of work of the MELQO project (UNESCO, 2017). Across these initiatives and others, ECD measurement analyses have advanced to cross-country invariance analyses on some existing measures. Thus some information is emerging both on the psychometric structure of multi-domain assessments of child development *within* countries, as well as whether these measures function well *across* countries in order to permit comparisons. Such information could be used in the process of creating a new "criterion" scale across the full range from birth to age 5 or 6 that would advance the field towards testing a single measure.

Two examples can illustrate the efforts to create a single criterion scale. WHO is leading a consortium to create a single criterion scale for 0-3 year olds. UNICEF, with its infrastructure for conducting multiple nationally representative samples in a sustained manner across years, could adapt and extend its ECDI measure (thus far for 3-6 year olds) with input from the initiatives of the past decade that have led the field towards cross-country comparable measures. Such an effort could integrate the WHO scale on some constructs that may be suitable for measurement across the 0-3 and 3-6 year old age ranges. The advantages of this process include the leveraging of resources for large samples at the country level for many countries that may not otherwise currently have these resources; experience in a range of regions; and the large number of LMICs within which the MICS is currently fielded.

We note that the process of creating a single criterion measure could also benefit from the experience of the creation and revision of the PISA cross-national assessments or others such as TIMSS, PIRLS, etc. (and currently the supplementation of the PISA with the PISA for Development, aimed for LMIC use). This is particularly relevant to the extension to rich countries required in any development of a new criterion measure. In the PISA development process, for example, initial wide-ranging expert consensus on a measurement framework at the level of constructs occurred, followed by convening panels of experts to develop items in specific domains; phases of pre-piloting in multiple countries with relatively small samples to ascertain meaning of items and variation in response to tasks, assessors, and administration formats; more systematic piloting of items across countries; item revision and selection for large-scale piloting; and finally nationally representative administration across countries (OECD, 2000a, 2000b). Many of these steps have been carried out in the ECD work of the MELQO initiative and others, but not all.

However, there are challenges facing such an effort as well, which should be taken into account. They include:

1) The need to continue to support country-level adaptation processes. For country-level use, the stakeholder process to build consensus towards national measurement of early childhood assessment for monitoring purposes and to inform policies in areas such as quality improvement and teaching and learning can and should be comprehensive. A single criterion measure can be used but could also be supplemented, for example, in particular countries with culturally specific constructs of child development that are relevant to goals for children's learning, behavior and development. Some countries may choose to use their own measures, and this is supported within the SDG process.
2) The definition of "on track" and "off track." No current widely used early childhood development measure among those mentioned in this document has established cutoffs for on vs off track. This is in part because these are not designed as screening measures. However, the development of national norms can be done without expectation of a cross-country, uniform definition of on and off track. Technical work to establish a consensus on this process within and across countries is necessary in the field of ECD.
3) The unprecedented range of country contexts. None of the current initiatives or existing measures in the field of ECD have been widely administered or analyzed across both LMIC and rich-country contexts. Should a single measure be developed from the basis of the UNICEF ECDI and other existing measures with cross-country data, it will be vital to consider cross-country measures that have been fielded in rich countries, including current initiatives of the OECD, the EU, and other entities. Some of the ECD measures recently fielded in multiple LMICs are starting to be applied in rich countries; these learnings should also be integrated.
4) Need to include both caregiver / adult report and direct child assessment. A consensus is building in the ECD field that measurement of some domains of development benefit

from the integration of information from adults who spend substantial time with children (caregivers / parents; teachers) and direct child assessment. For example, adults familiar with children's behaviors in home and/or care settings may be in a better position to observe low-frequency behaviors such as aggression than can be assessed in an assessor administered task. Conversely, direct assessments may be more appropriate when certain skills are not ones that adults in children's lives are used to noticing, but may nevertheless be predictive of later outcomes (e.g., aspects of executive function), or when complex skills benefit from standard stimuli (e.g., comprehension of a sentence). It is undeniable that direct child assessment is more costly, with training to reliability more difficult at scale than with more straightforward survey-based measures. However, options such as random subsamples, within a larger nationally representative sample, for direct child assessment modules should be considered. This approach may reduce the overall costs of adding a direct child assessment portion to an adult-reported measure in a nationally representative sample.

5) A measure that vertically equates some domains of development across wider age ranges than 0-3 and 3-6. The vertical equating required to achieve measures of some areas of development that are meaningful to measure across the full age span is very challenging, given the rapidity of development in these years and the qualitative changes in skills that occur, not just quantitative. Yet existing measures collected across countries have for the most part been restricted to the 0-3 vs the 3-6 year old age range. The integration of the WHO consortium on 0-3 measurement and current efforts in the 3-6 year olds age range would be critical for this effort. It is likely that only some constructs of development are suited to integration across 0-3 and 3-6.
6) Alignment with later learning targets and indicators in SDG 4. The continuum of learning and development stretches from birth to adulthood. Alignment of 4.2.1 with indicator 4.1.1, in particular (and especially the grade 2 or 3 indicator), is important to enable nations to track how learning unfolds in the first 8 years of life.
7) Alignment with other SDG targets and indicators. The alignment of SDG 4.2.1 with other goals, targets and indicators in the areas of health, mental health, nutrition, and child protection is signaled in the wording of 4.2.1, which (unlike the primary schooling indicators) integrates health and psychosocial well-being. Such alignment can move beyond health and psychosocial well-being to consider relationships with other SDG indicators outside of Goal 4.
8) Integration of member nation input into development of a criterion measure. The input of UN member nations into the development of the SDGs, including Target 4.2., was unprecedented in history. Continued input into the development of a criterion measure for 4.2.1 is vital for ultimate use of the measure at the country level and the global processes to track SDG 4.

B. Assessing our options: Which measurement strategy maximizes both validity and use?

A single organizing question to help in assessing these options might be: What approach maximizes validity, feasibility and productive national and cross-national use to inform policy and practice? To answer such a question, agreement on the meaning and evidence to support validity must be a starting point.

<u>What does "validity" mean in cross-national measurement of early childhood development?</u>
A central goal of assessment in the fields of child development and education is to achieve measurement with evidence of validity. Current notions of validity consider it a unitary construct supported by evidence in the context of use. With the new demand for policy-relevant data, the types of evidence that should be weighed in assessing validity include but go beyond older conceptualizations of validity (for example, the trio of content, criterion-related, and construct validity and their subtypes; Cronbach & Meehl, 1955).

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999).

Note that this overall single conception of validity supercedes the subtypes of content, criterion-related and construct validity. Five kinds of evidence have been put forward to support validity according to this definition (Goodwin & Leech, 2003). These include many of the traditional subtypes of content, criterion-related and construct validity.

First, **evidence based on test content** requires that expert consensus be achieved on the match between item and task content and the construct that is being measured, and whether the content reflects bias or differential match between content and construct for particular groups (e.g., as defined by gender, language, culture, etc.). In order to maximize this kind of evidence, the "group consensus" approach in test development and refinement, as well as testing explicitly for bias through qualitative research, cognitive testing, understanding of variation in experience of the testing context, analyses to test for measurement invariance across diverse populations and other methods can be employed.

Second, **evidence based on responses of test takers** examines potential unintended responses such as those based on social desirability, unfamiliarity with the administration approach or testing context. This form of evidence in the case of measures of Indicator 4.2.1 includes, for example, analyzing reasons for non-response that go beyond lack of underlying ability, to discomfort, anxiety, or response to the assessor and setting.

Third, **evidence on internal structure** taps whether the components of a measure adequately reflect the subdomains of a construct. In the case of the multi-domain measures of early childhood development, this includes whether the instrument adequately reflects physical, learning, and psychosocial domains. The learning domain is often considered to include potential subdomains such as language / early literacy; numeracy, spatial and quantitative skills; and executive function or approaches to learning. This kind of evidence is most often analyzed through exploratory and confirmatory factor analyses.

Fourth, **evidence based on relations to other measures** includes traditional forms of criterion-related and construct validity, such as concurrent, predictive, convergent, and discriminant validity. Such evidence also includes the important criterion of sensitivity to intervention. This is particularly important in the SDG 4.2 context, as the overall target links early childhood development to the quality of policies and programs that support it.

Finally, **evidence based on the consequences of testing** focuses centrally on uses of the assessment. The ability of a measure to inform practice and policy involves not only feasibility of administration at large scale with regular periodicity, but the links to practice and policy decision-making. In this regard, the MELQO initiative, building on prior efforts, proposed that both measurement of quality of early learning environments and measurement of early childhood development outcomes was necessary to most powerfully inform policy and practice (UNESCO, 2017).

C. Next Steps and Questions to Guide Discussion

The SDGs provide a unique opportunity for building a global ECD measurement strategy that significantly enhances the reliability, feasibility and comparability of existing ECD data. Moving forward on any of the strategies outlined above will require a greater degree of systematic data collection, coordination among measures developers and experts, and input from stakeholders than has taken place in the past. Below we have outlined options for pursuing consensus on these strategies:

- Next Step 1. Building on prior consensus work, begin by agreeing on a general conceptual framework (e.g., of early childhood development domains) to guide measurement. This will serve as the groundwork for the next steps of clarifying what could be measured across countries, and where existing data may be available to help inform decisions on constructs and items. Several such consensus meetings have occurred recently; however, there are still areas of lack of consensus such as some of the eight areas of challenges listed above.

- Next Step 2. Address questions of validity in light of the strong policy emphasis of SDG-related data, the unique aspects of early childhood development relative to later phases of learning, and the need to ensure equity and cultural relevance across a range of countries. A convening on technical standards and guidelines for use, synthesis and harmonization of data from existing measures; standards for cross-country comparability of the option of a single criterion measure; and the criteria for dimensions of validity in such an effort is critical. Engaging a wide range of psychometricians, researchers and other expert stakeholders who can help to assess the feasibility of each approach, including the cost and coordination requirements for pursuing each of the options, is required. No convening to date, for example, has brought together the small group of psychometric experts who have worked on cross-country analyses of existing ECD measures. Building a technical consensus for the next phase of work would be an important step in the next advances towards global measurement of SDG Indicator 4.2.1.

Questions to Guide Discussion:

1) Among Options 1, 2 and 3, which seem feasible in the short run (next 12-18 months)? In the medium run (1.5-3 years)? In the longer run (3-5 years)?
2) What is the best plan for making progress on the three Next Steps noted immediately above in this section? Are any important Next Steps missing?

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, K., & Raikes, A. (2017). *Key measurement questions for Indicator 4.2.1 (Discussion paper for GAML Taskforce 4.2)*.

Bornstein, M. H., Britto, P. R., Nonoyama-Tarumi, Y., Ota, Y., Petrovic, O., & Putnick, D. L. (2012). Child development in developing countries: introduction and methods. *Child Development*, *83*(1), 16-31.

Chan, K. S., Gross, A. L., Pezzin, L. E., Brandt, J., & Kasper, J. D. (2015). Harmonizing Measures of Cognitive Performance Across International Surveys of Aging Using Item Response Theory. *Journal of aging and health*, *27*(8), 1392-1414.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Gatz, M., Reynolds, C. A., Finkel, D., Hahn, C. J., Zhou, Y., & Zavala, C. (2015). Data harmonization in aging research: Not so fast. *Experimental aging research*, *41*(5), 475-495.

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH toolbox for assessment of neurological and behavioral function. *Neurology*, *80*(11 Supplement 3), S2-S6.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing. *Measurement and Evaluation in Counseling and Development*, *36*(3), 181-192.

Kagan, S. L., & Britto, P. R. (2005). *Going global with indicators of childdevelopment.* New York: UNICEF.

Jacobusse, G., Van Buuren, S., & Verkerk, P. H. (2006). An interval scale for development of children aged 0–2 years. *Statistics in medicine*, *25*(13), 2272-2283.

Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science, 39*(1), 1-22.

McCoy, D. C., Peet, E. D., Ezzati, M., Danaei, G., Black, M. M., Sudfeld, C. R., ... & Fink, G. (2016). Early childhood developmental status in low-and middle-income countries: national, regional, and global prevalence estimates using predictive modeling. *PLoS Med*, *13*(6), e1002034.

McCoy, D. C., Sudfeld, C. R., Bellinger, D. C., Muhihi, A., Ashery, G., Weary, T. E., ... & Fink, G. (2017). Development and validation of an early childhood development scale for use in low-resourced settings. *Population health metrics*, *15*(1), 3-.

OECD (2000a). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy.* Paris: Author.

OECD (2000b). *PISA 2000: Technical Report.* Paris: Author.

Raikes, A., & Anderson, K.L. (2017). *Key measurement questions for SDG 4.2.1 (discussion paper).* Montreal: UNESCO Institute for Statistics, Global Alliance to Monitor Learning, Target 4.2 Task Force.

Rao, N., Sun, J., Ng, M., Becher, Y., Lee, D., Ip, P., & Bacon-Shone, J. (2014). Validation, Finalization and Adoption of the East Asia-Pacific Early Child Development Scales (EAP-ECDS). UNICEF, East and Pacific Regional Office.

Ravens-Sieberer, U., Erhart, M., Rajmil, L., Herdman, M., Auquier, P., Bruil, J., ... & Mazur, J. (2010). Reliability, construct and criterion validity of the KIDSCREEN-10 score: a short measure for children and adolescents' well-being and health-related quality of life. *Quality of Life Research*, *19*(10), 1487-1500.

Reardon, S.F., Kalogrides, D., & Ho, A.D. (2016). *Linking U.S. school district test core distributions to a common scale, 2009-2013*. Palo Alto, CA: Stanford Center for Education Policy Analysis.

Schmitt, J., Apfelbacher, C., Spuls, P. I., Thomas, K. S., Simpson, E. L., Furue, M., ... & Williams, H. C. (2015). The Harmonizing Outcome Measures for Eczema (HOME) roadmap: a methodological framework to develop core sets of outcome measurements in dermatology. *Journal of Investigative Dermatology*, *135*(1), 24-30.

UN Statistical Commission (2016). *Provisional proposed tiers for global SDG indicators*. New York: Author.

UN Statistical Commission (2017). *Revised list of global Sustainable Development Goal indicators.* New York: Author.

UNESCO (2017). *Overview: Measuring Early Learning and Quality Outcomes (MELQO).* Paris: Author.

Verdisco, A., Cueto, S., & Thompson, J. (2016). *Early Childhood Development: Wealth, the Nurturing Environment and Inequality First Results from the PRIDI Database*. Inter-American Development Bank.

Williamson, P. R., Altman, D. G., Blazeby, J. M., Clarke, M., Devane, D., Gargon, E., & Tugwell, P. (2012). Developing core outcome sets for clinical trials: issues to consider. *Trials*, *13*(1), 132.

Wolf, S., Halpin, P., Yoshikawa, H., Pisani, L., Dowd, A.J., & Borisova, I. (2017). *Assessing the construct validity of Save the Children's International Development and Early Learning Assessment (IDELA).* Manuscript under review.